AD-A103 956 ROME AIR DEVELOPMENT CENTER GRIFFISS AFB NY F/6 12/1
AN ERROR ANALYSIS TECHNIQUE FOR STATISTICAL HYPOTHESIS TESTING.--ETC(U)
MAR 81 J F LENNON, R J PAPA
UNCLASSIFIED RADC-TR-81-81 NL

1 OF 1
AD A
103956

END
DATE
FILMED
10-81
DTIC

AD A103956

RADC-TR-81-81
In-House Report
March 1981

AN ERROR ANALYSIS TECHNIQUE FOR STATISTICAL HYPOTHESIS TESTING
Generation of Jointly Distributed Random Variates

John F. Lennon
Robert J. Papa

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-81-81 has been reviewed and is approved for publication.

APPROVED: Philipp Blacksmith

PHILIPP BLACKSMITH
Chief, Electromagnetic Techniques Branch
Electromagnetic Sciences Division

APPROVED: Michael J. O'Brien

MICHAEL J. O'BRIEN, Major, USAF
Acting Chief, Electromagnetic Sciences Division

FOR THE COMMANDER: John P. Huss

JOHN P. HUSS
Acting Chief, Plans Office

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-81-81	2. GOVT ACCESSION NO. AD-A102 956	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) AN ERROR ANALYSIS TECHNIQUE FOR STATISTICAL HYPOTHESIS TESTING Generation of Jointly Distributed Random Variates		5. TYPE OF REPORT & PERIOD COVERED In House 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) John F. Lennon Robert J. Papa		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Deputy for Electronic Technology (EEC) Hanscom Air Force Base Massachusetts 01731		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F, 2305J407
11. CONTROLLING OFFICE NAME AND ADDRESS Deputy for Electronic Technology (EEC) Hanscom Air Force Base Massachusetts 01731		12. REPORT DATE March 1981 13. NUMBER OF PAGES 40
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Statistics
Hypothesis testing
Random variate generation
Error probabilities

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

As a part of the study of electromagnetic scattering from rough terrain, various geographical sites were characterized by applying statistical analysis techniques to digitized terrain data bases. One aspect was the use of decision theory formulation to assign an appropriate distribution function to the surface height variations in the regions. This report discusses the decision processes used and the errors to be expected. There are two main themes: assessment of statistical hypothesis testing and application of a numerical technique to

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (Continued)

determine the errors inherent in the decisions. There are a number of assumptions and constraints that affect the form of the hypothesis test developed for the terrain study and these will be considered. The numerical technique consists in generating jointly distributed random variates in a Monte Carlo type computer procedure. Its application to the specification of decision errors will be described. The formulation allows extension of standard analytic error analysis techniques to cases where the analysis would be intractable. Results are presented and the agreement with analysis for various cases is shown.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Preface

The authors gratefully acknowledge the contributions of Dr. John K. Schindler for many helpful comments, criticisms, and discussions and Ms. Nancy Kerwin and colleagues (ARCON) who provided extensive programming support.

Contents

Tables

An Error Analysis Technique for Statistical Hypothesis Testing

Generation of Jointly Distributed Random Variates

1. INTRODUCTION

There are two main themes in this report. The first is an assessment of the type of hypothesis test used in our terrain characterization studies.[1,2] The second is the construction of a numerical technique to allow us to carry out that assessment. Computer generated, jointly distributed sets of random variates are used in a Monte Carlo formulation that verifies the decision errors to be expected for various hypothesis tests. There is a need for this type of approach since appropriate analytic results cannot always be obtained.

In the initial characterization study, the problem of interest is that of selecting appropriate statistical quantities to describe a large terrain region that is considered to be made up of smaller subareas (~4 km^2). The main topographic feature is the distribution of heights within these subregions. This places some constraints on the form of the particular statistical analysis carried out for the report. The eventual goal of these studies is the development of mathematical descriptions of

---

(Received for publication 20 March 1981)

1. Lennon, J.F. and Papa, R.J. (1980) Statistical Characterization of Rough Terrain, RADC-TR-80-9, RADC/EE Hanscom AFB, MA.
2. Papa, R.J., Lennon, J.F., and Taylor, R.L. (1980) Electromagnetic Wave Scattering From Rough Terrain, RADC-TR-80-300, RADC/EE Hanscom AFB, MA.

each of the subareas for use in calculation of the scattering of electromagnetic waves from the uneven terrain surface. Each region is characterized by a geologic code and several statistical parameters. In particular, we are concerned with being able to associate a probability density function (PDF) with the range of heights in the subregions and to determine parameters that make the general PDF explicit.

When we start from some observed values, the most general situation would be that neither the form of the PDF, nor the parameters which enter into the expression for the PDF are known. In order to obtain some information on these parameters, estimation theory is used.[3] Then, the parameters are incorporated into a PDF, the form of which is not known a priori and must be determined. An hypothesis testing procedure is developed to accomplish this.

The particular hypothesis testing procedure selected in the present case allows only a binary decision process to be considered. Hence, the discrimination is restricted to two forms of the PDF. The test is based upon the maximum a posteriori probability criterion.[4] This is equivalent to the minimum error probability criterion. The procedure for hypothesis testing may be applied quite generally. Indeed, where the number of cases allow, a better test would be to see which density would be more likely to have generated several realizations rather than the single one used here.

For whatever type of PDFs and hypothesis tests employed in the characterization, there is still the concern as to the errors involved in the decision process. In using the results in various electromagnetic calculations of multipath and clutter, it is desirable to be able to assess the reasonableness of the terrain feature distributions which have been decided upon, and to understand the implications of the decisions.

In the form of hypothesis test used in the present studies, the decision is between simple alternatives. The respective probabilities are written as a quotient and the decision statistic, a function of the random variates, is determined. The decision is based on the location of the statistic value with respect to the two regions of the decision space assigned to the respective hypotheses. In the case of simple alternatives, the two regions, although not necessarily connected, do cover the entire decision space. The concepts of the decisions and hypothesis testing in general will be discussed more fully in Section 4.

One aspect of this topic that should be kept in mind is that this particular test is concerned with minimizing not the errors of each alternative hypothesis but only

---

3. Jenkins, G.M., and Watts, D.G. (1968) Spectral Analysis and Its Applications, Holden-Day, San Francisco, CA.

4. Whalen, A.D. (1971) Detection of Signals in Noise, Academic Press, New York, NY.

the total error of an incorrect decision where both possibilities are taken into account simultaneously. Thus, based on the forms of the alternative densities, it may occur that the hypothesis test would allow decision regions that produce relatively high errors in one alternative, so long as the total error is minimized. If this is not acceptable, other forms of the test will have to be devised.

The actual assessment of the errors involved in the decision sometimes can be made analytically. One such case occurs when the function of the random variates that represents the decision statistic is sufficiently simple to allow calculation of its cumulative probability over the various decision regions. In general, though, when we are concerned with sequences of random variates, dissimilar PDFs, and probabilities that are not independent, the statistic is often intractable analytically. In such instances an alternative approach to assessing the possible decision errors has to be employed. Since we are concerned with a statistical process in the sense of determining that a specific set of observables $(Z_1, Z_2, \ldots, Z_N)$ of the set of jointly distributed variates $(z_1, z_2, \ldots, z_N)$ is from a particular distribution, then one possibility would be to approach the problem from a Monte Carlo point of view.[5] To implement that approach, we use a computer model to generate sets of jointly distributed random variates from selected probability densities. Next, the corresponding form of the hypothesis test is applied to a large number of these independent observations of the variates to determine the probability of an incorrect decision. The probability is modeled on the basis that there are a given number of incorrect decisions by the test in some large number of tries. The key to this approach, then, is to be able to generate the desired sets of variates having the given joint probability density functions. This is relatively straightforward when the variates are independent, but can be quite complex when they are not.

The report first addresses this problem of producing the sets of variates. Then, specific cases are considered. Some aspects of the hypothesis testing are treated. The results of the Monte Carlo approach in evaluating the tests are then compared with some analytic results. Finally, the specific application to terrain height analysis is developed and the implications of the errors are discussed.

2. GENERATION OF NONUNIFORM RANDOM VARIATES

A standard tool in statistical theory is the generation of psuedorandom numbers based on the uniform probability distribution.[5] Variates representing other distributions can also be determined. In this section we will discuss several aspects of the general theory. Both single variate and multivariate forms will be considered. Independent and dependent distributions will be examined separately.

---

5. Shreider, Yu., A. (1966) The Monte Carlo Method, Pergamon Press, Oxford.

2.1 Relation to Uniform Variates

In the [illegible], there [illegible] numbers, Z_n [illegible] distribution [illegible] $P($[illegible] numbers U_n which are uniformly distributed [illegible] relation may be expressed as follows

$$U_n = \int_{-\infty}^{Z_n} p(z)\, dz.$$

where

$$0 \le U_n \le 1 \quad \text{and } n = 1, 2, 3, \ldots$$

A detailed discussion of this relation is found in Shreider.[5]

The procedure used to generate random variates, Z_n having the probability density $p(z)$ is based upon the above relation. First, standard computer algorithms are used to generate a set of uniformly distributed random variates, U_n. Next, a probability density function $p(z)$ is selected. Then, in the integrals

$$F_n(Z_n) = \int_{-\infty}^{Z_n} p(z)\, dz$$

upper limits, Z_n are determined such that $F_n = U_n$.

If $F(Z_n)$ can be formed analytically in terms of known functions, then Z_n can be determined directly by finding the inverse relation such that $Z_n = F^{-1}(U_n)$. Examples of this case include the Gaussian density,

$$p_G(z) = (2\pi\sigma^2)^{-1/2} \exp[-z^2/2\sigma^2]$$

and the Laplacian density,

$$p_L(z) = \left(\frac{\alpha}{2}\right) \exp[-\alpha|z|].$$

Alternatively, if $F_n(Z_n)$ has to be evaluated numerically (using some quadrature formula, such as Simpson's rule) then the value of Z_n such that $F_n(Z_n) = U_n$ is obtained by iteration.

So far, we have been discussing the relation between sets of uniform random numbers individually related to corresponding sets satisfying a different probability distribution. Next, we address the question of the relation between jointly distributed sets of random variates and equivalently sized groups of uniform random numbers.

2.2 Jointly Distributed Random Variates

In order to consider this case, we look at the relation between the joint density function and its associated conditional densities,[6,7]

$$
\begin{aligned}
p(z_1, z_2, \ldots, z_N) &= p(z_N \mid z_{N-1}, \ldots, z_1)\, p(z_{N-1}, \ldots, z_1) \\
&= p(z_N \mid z_{N-1}, \ldots, z_1)\, p(z_{N-1} \mid z_{N-2}, \ldots, z_1) \cdots \\
&\qquad \cdots\, p(z_2 \mid z_1)\, p(z_1).
\end{aligned}
$$

We use these relations and the definitions relating to cumulative distribution functions and consider, just as for the single random variate case, a set of jointly distributed variates $(Z_1, Z_2, \ldots, Z_N)$ that occupy a volume $\Delta V_z = \Delta z_1 \Delta z_2 \ldots \Delta z_N$ in a joint probability space. These variates are related to a corresponding set of uniformly distributed random variates $(U_1, U_2, \ldots, U_N)$ having the associated volume $\Delta V_u = \Delta u_1 \Delta u_2 \ldots \Delta u_N$. The distinction here is that we now work with successive single-variate, conditional densities each of which depends on the previously determined elements of the set. To illustrate the specifics of the procedure, we will first outline the two variate case[5] and then proceed to the general multivariate case.

For two variates we have,

$$p(z_1, z_2) = p(z_2 \mid z_1)\, p(z_1) .$$

Here $p(z_1)$ is the marginal density given by

$$p(z_1) = \int_{-\infty}^{\infty} p(z_1, z_2)\, dz_2 .$$

6. Papoulis, A. (1965) Probability, Random Variables and Stochastic Processes, McGraw-Hill, New York, NY.
7. Mood, A.M., and Graybill, F.A. (1963) Introduction to the Theory of Statistics, McGraw-Hill, New York, NY.

For a given $p(z_1, z_2)$ we first use the technique of Section 2.1 to get Z_1 as the realization of $p(z_1)$ based on some U_1,

$$U_1 = \int_{-\infty}^{Z_1} p(z_1)\, dz_1 .$$

Next, the resultant Z_1 is substituted into the conditional probability density $p(z_2 | z_1)$ and the integral

$$F(Z_1, \xi_2) = \int_{-\infty}^{\xi_2} p\,(z_2 | Z_1)\, dz_2$$

is constructed. Now this integral is the same form as the integrals considered in Section 2.1, so that a second random variate Z_2 may be generated, which belongs to the density $p(z_2 | z_1)$. This is accomplished by using the relation

$$U_2 = \int_{-\infty}^{Z_2} p(z_2 | Z_1)\, dz_2 .$$

The two random variates, Z_1 and Z_2, satisfying respectively the univariate marginal and the related conditional density, jointly have the density $p(z_1, z_2)$. The procedure can be repeated as often as desired to obtain multiple realizations, $(Z_1, Z_2)_n$, of the joint distribution.

For the multivariate case we use the univariate marginal, the successive conditional densities, and the corresponding marginal densities,

$$p(z_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(z_1, z_2, \ldots, z_N)\, dz_2, dz_3 \ldots dz_N;$$

$$p(z_2 | Z_1) = p(z_2, Z_1)/p(Z_1);$$

$$p(z_3 | Z_1, Z_2) = p(z_3, Z_2, Z_1)/p(Z_1, Z_2);$$

$$\vdots$$

$$p(z_N | Z_1, Z_2, \ldots, Z_{N-1}) = p(z_N, Z_{N-1}, \ldots, Z_1)/p(Z_1, Z_2, \ldots, Z_{N-1}) .$$

The set of variates having the joint probability density $p(z_1, z_2, \ldots, z_N)$ is then obtained by the successive integrations

$$U_1 = \int_{-\infty}^{Z_1} p(z_1)\, dz_1 ;$$

$$U_2 = \int_{-\infty}^{Z_2} p(z_2 | Z_1)\, dz_2 ;$$

$$\vdots$$

$$U_N = \int_{-\infty}^{Z_N} p(z_N | Z_1, Z_2, \ldots, Z_{N-1})\, dz_N .$$

Again, the sequence can be repeated as often as desired for multiple realizations $(Z_1, Z_2, \ldots, Z_N)_n$.

2.3 Independent Joint Distributions

The random variates in the joint density function, $p(z_1, z_2, \ldots, z_N)$ are independent if and only if

$$p(z_1, z_2, \ldots, z_N) = p_1(z_1)\, p_2(z_2) \ldots p_N(z_N)$$

for all values for which the random variates are defined. Similarly, the cumulative distribution of N independent random variates is given by

$$F(\xi_1, \xi_2, \ldots, \xi_N) = \int_{-\infty}^{\xi_1} dz_1 \int_{-\infty}^{\xi_2} dz_2 \ldots \int_{-\infty}^{\xi_N} dz_N\, p_1(z_1)\, p_2(z_2) \ldots p_N(z_N)$$

$$= F_1(\xi_1)\, F_2(\xi_2) \ldots F_N(\xi_N)$$

where

$$F_m(\xi_m) = \int_{-\infty}^{\xi_m} dz_m \, p_m(z_m) .$$

Thus, when we have the case of independent variates where the probability density functions are the same for all the variates, the multivariate random number procedure is simplified. It reduces to the situation where all that is required is successive implementations of the single variate determination. We then have a set $(Z_1, Z_2, \ldots, Z_N)$ of independent variates each satisfying the corresponding probability density and, as before, additional sets can be generated.

3. EXAMPLES

In the preceding section we have discussed the general technique for producing sets of random variates that satisfy given joint probability density relations. Both the simplified procedure for the case where the variations are independent and the more complicated form when they are not independent were outlined. In the various analyses treated in this report several different density functions were used and these examples will be discussed in this section.

3.1 Independent Variates

For the case where the jointly distributed variates are independent, the only form required is that of the single variate marginal density. Two types are used, Gaussian and Laplace. For all the generating routines, the variates are taken to satisfy zero mean, unit variance distributions. The implications of this assumption will be discussed in Section 7.

The first case is the form for the multivariate Gaussian density:

$$p_G(z_1, z_2, \ldots, z_N) = (2\pi)^{-N/2} \exp[-1/2\,(z_1^2 + z_2^2 + \ldots + z_N^2)] .$$

The corresponding form for the multivariate Laplace density is

$$p_L(z_1, z_2, \ldots, z_N) = (\sqrt{2})^{-N} \exp[-\sqrt{2}\,(|z_1| + |z_2| + \ldots + |z_N|)] .$$

These two forms are the only two cases of independent variates that were generated in the course of the study. All the other results are for general distributions where the variates are not independent.

3.2 Nonindependent Variates

In this category, the fact that the joint densities cannot be written as products of single variate independent probabilities means that the successive marginal and conditional densities must be included in the random number generation (RNG) process. It should be noted that in these cases we will be dealing with uncorrelated forms of the variates where the uncorrelated variates are not necessarily independent. Three specific examples were considered in various phases of the overall study. For each of these cases we will indicate the general N-variate form of the PDF and the associated general marginal density relation.

The first form is that of the particular version of the multivariate exponential density used in our initial terrain analysis:[1]

$$p_E(z_1, \dots, z_n) = \left(\frac{(N+1)^{N/2}}{2^N \pi^{\frac{N-1}{2}} \Gamma\left(\frac{N+1}{2}\right)} \right) \exp[-\sqrt{N+1}\,(z_1^2 + z_2^2 + \dots + z_N^2)^{1/2}].$$

Its general L-variate marginal density is,

$$p_E(z_1, z_2, \dots, z_L) = \left(\frac{2}{\pi^{L/2}\, \Gamma\left(\frac{N+1}{2}\right)} \right) \left(\frac{N+1}{4}\right)^{\frac{N+L+1}{4}}$$

$$(z_1^2 + z_2^2 + \dots + z_L^2)^{\frac{N-L+1}{4}} \times$$

$$\times\, K_{\frac{N-L+1}{2}}\left(\sqrt{L+1}\sqrt{z_1^2 + z_2^2 + \dots + z_L^2} \right) \qquad \text{for } L \le N$$

where $\Gamma(u)$ is the Gamma function and $K_y(X)$ is the modified Bessel function of the second kind. This form and those of the following two cases have been described previously.[8]

8. Lennon, J. F. (1980) The Derivation of a Multivariate Probability Density Function Having an Exponential-Type Bivariate Marginal Density, RADC-TR-80-153, RADC/EE, Hanscom AFB, MA.

The second PDF that is of interest is the form of Bessel function density that has a bivariate exponential marginal density:[8]

$$p_K(z_1, \ldots, z_N) = \left(\frac{3^{N/2}}{2^{\frac{N-1}{2}} \pi^{\frac{N+1}{2}}}\right) \left(\sqrt{3}\sqrt{z_1^2 + z_2^2 + \ldots + z_N^2}\right)^{-\left(\frac{N-3}{2}\right)} \times$$

$$\times K_{\frac{N-3}{2}}\left(\sqrt{3}\sqrt{z_1^2 + z_2^2 + \ldots + z_N^2}\right).$$

For this density the L-variate marginal form is:

$$p_K(z_1, \ldots, z_L) = \left(\frac{3^{L/2}}{2^{\frac{L-1}{2}} \pi^{\frac{L+1}{2}}}\right) \left(\sqrt{3}\sqrt{z_1^2 + z_2^2 + \ldots + z_L^2}\right)^{-\left(\frac{L-3}{2}\right)} \times$$

$$\times K_{\frac{L-3}{2}}\left(\sqrt{3}\sqrt{z_1^2 + z_2^2 + \ldots + z_L^2}\right).$$

The final example of nonindependent variates, though, was developed specifically to verify some elements of this analysis and the discussion of the determination of the appropriate normalization factors for the PDF will be presented in Appendix A.

This final form resembles a Gamma type distribution and will be referred to in those terms in following sections:

$$p_\Gamma(z_1, \ldots, z_N) = C_1(|z_1| + \ldots + |z_N|) \exp[-C_2(|z_1| + \ldots + |z_N|)]$$

where

$$C_1 = C_2^{N+1}/(N2^N) \quad \text{and} \quad C_2 = \sqrt{2(N+2)/N}.$$

The marginal density form is

$$p_\Gamma(z_1, \ldots, z_L) = \left(\frac{2^{N-L} C_1}{C_2^{N-L+1}}\right) [(N-L) + Y] \exp(-Y) \qquad \text{for } L \le N$$

where

$$Y \equiv [C_2(|z_1| + \ldots + |z_N|)] \; .$$

3.3 Illustration of the Procedure

In Section 2, the overall approach to generating random numbers having specified probability densities was described. This section has presented the particular types of PDF. We will now proceed to outline the technique used to obtain the desired sets of variates. In the two examples we will use $N = 5$ as the number of jointly distributed variates.

For both independent and nonindependent cases the starting point is the same. We use standard computer techniques to generate variates satisfying the uniform distribution in the range 0 to 1. The computer algorithms produce a sequence of psuedorandom numbers, α_i, that have weak statistical correlation, the PDF of the psuedorandom numbers approximates an uniform PDF, and the program is stable, that is, the PDF remains the same for all output variates. For the actual program used, the algorithm is,

$$\alpha_{j+1} = 2^{31} \beta_{j+1} ; \qquad \beta_o = 466364003.0$$

and

$$\beta_{j+1} = 16807 \, \beta_j \; (\mathrm{MOD} \; 2^{31}-1) \; .$$

The resultant sequences are then transformed into the desired variates.

For the independent Gaussian case we just apply the relation

$$U_i = \int_{-\infty}^{Z_i} (2\pi)^{-1/2} \; \exp\,(-z_i^2/2)\, dz_i \qquad i = 1, 2, \ldots, 5$$

five successive times to form the set of desired jointly distributed normal variates, $\{z_i\}$. The evaluation is based on standard representations of the inverse error function for the required ranges of z_i.

If we use our form of exponential PDF as the example for the nonindependent case, the sequence of five numerical integrations required to determine the jointly distributed set of variates is,

$$U_1 = \int_{-\infty}^{Z_1} (3\sqrt{6}/16)(1+\sqrt{6}|z_1| + 2z_1^2) \exp\left(-\sqrt{6z_1^2}\right)dz_1 ,$$

$$U_2 = \int_{-\infty}^{Z_2} (p(Z_1))^{-1} [(9/4\pi)(Z_1^2+z_2^2)] K_2\left(\sqrt{6}\sqrt{Z_1^2+z_2^2}\right)dz_2 ,$$

$$U_3 = \int_{-\infty}^{Z_3} (p(Z_1, Z_2))^{-1}\left[(3\sqrt{6}/16\pi)\left(1+\sqrt{6}\sqrt{Z_1^2+Z_2^2+z_3^2}\right)\right] \times$$

$$\times \exp\left[-\sqrt{6}\sqrt{Z_1^2+Z_2^2+z_3^2}\right]dz_3 ,$$

$$U_4 = \int_{-\infty}^{Z_4} (p(Z_1, Z_2, Z_3))^{-1}\left[(9\sqrt{6}/8\pi^2)\left(\sqrt{Z_1^2+Z_2^2+Z_3^2+z_4^2}\right)\right] \times$$

$$\times K_1\left(\sqrt{6}\sqrt{Z_1^2+Z_2^2+Z_3^2+z_4^2}\right)dz_4 ,$$

and

$$U_5 = \int_{-\infty}^{Z_5} (p(Z_1, Z_2, Z_3, Z_4))^{-1}(9\sqrt{6}/16\pi^2) \times$$

$$\times \exp\left[-\sqrt{6}\sqrt{Z_1^2+Z_2^2+Z_3^2+Z_4^2+z_5^2}\right]dz_5 .$$

The sequence $\{Z_1, Z_2, Z_3, Z_4, Z_5\}$ obtained from the successive numerical integrations now forms the set of jointly distributed random variates. It should be noted that not all cases require numerical integration; for some PDFs the integration may be analytic.

4. HYPOTHESIS TESTING

As explained in Morrison,[9] statistical inference may be divided into two general categories. The first category is concerned with the estimation of distribution functions, the parameters of such functions when their mathematical form is specified, or the parameters of models related to random variables. The second category is devoted to the problem of testing the validity of hypotheses about distribution functions and their parameters. This section is concerned with statistical inference in the second sense. A further restriction is that only simple alternative hypothesis testing will be considered.

As was discussed in the report by Lennon and Papa,[1] the concern for our particular case is to determine whether a set of data (for example, terrain heights) is better described by a Gaussian or by another probability density function (PDF). Let hypothesis H_1 correspond to the case where the data set is from a Gaussian distribution and hypothesis H_o correspond to the case where the data set is from some other distribution. Let Y denote the values assumed by the random variable from which the data set is assumed to originate. Then let $P(H_o|Y)$ be the probability that H_o is true given the observation Y and let $P(H_1|Y)$ be the probability that H_1 is true given the observation Y. The discussion of hypothesis testing follows Whalen.[4] Let $p_o(y) \equiv p(y|H_o)$ represent the probability describing the data, given that H_o is true and let $p_1(y) \equiv p(y|H_1)$ denote the probability for the data given that H_1 is true. The sample space of observations y may be divided into two regions R_o and R_1 such that if a sample point $Y_i \epsilon y$ falls into R_o then H_o is chosen (denoted as decision D_o) and if the sample point falls into R_1 then H_1 is chosen (denoted as decision D_1).

For both hypotheses, the assignment of regions and the associated decision process involve two possible errors, type I being the rejection of a true hypothesis and type II being accepting a false one. For example, defining the bounds of R_o neglects both $P(Y \epsilon R_1|H_o)$ and $P(Y \epsilon R_o|H_1)$.

In terms of decisions relating to the hypothesis test evaluations, we postulate the density a priori and generate large numbers of sample observations from that given density. Thus what we are concerned with is the type I error for the validation procedure. For instance, if we have H_o as true then the error is

$$P(D_1|H_o) = \int_{R_1} P_o(y)\, dy \ ,$$

9. Morrison, D. F. (1976) Multivariate Statistical Methods, McGraw-Hill, New York, NY.

and correspondingly if we select the density of H_1 to generate the collection of sample sets, then the error that we evaluate with this technique is

$$P(D_o|H_1) = \int_{R_o} p_1(y)\,dy \quad .$$

It should be noted that when the hypothesis test approach is being applied to data from unknown densities and a decision made as to the most appropriate hypothesis, then the two types of error must be considered for each observation, Y_i.

4.1 Background

In testing hypotheses, one aspect is assessing relative costs of correct and incorrect decisions. We define C_{ij} as the cost associated with choosing hypothesis H_i when hypothesis H_j is true. In this context, the approach (known as the Bayes strategy) is to formulate relations for the average cost or risk for a decision and then to minimize this average cost. The average cost for the decision procedure is

$$\bar{C} = P(H_o)[P(D_o|H_o)C_{oo} + P(D_1|H_o)C_{1o}]$$

$$+ P(H_1)[P(D_o|H_1)C_{o1} + P(D_1|H_1)C_{11}] \ .$$

By using the relations

$$P(H_1) = 1 - P(H_o)$$

$$P(D_1|H_1) = 1 - P(D_o|H_1)$$

$$P(D_1|H_o) = 1 - P(D_o|H_o)$$

$$P(D_o|H_o) = \int_{R_o} p_o(y)\,dy$$

$$P(D_o|H_1) = \int_{R_o} p_1(y)\,dy$$

the average cost may be expressed as

$$\bar{C} = P(H_o)C_{1o} + [1 - P(H_o)]C_{11} + \int_{R_o} \{[1 - P(H_o)](C_{o1} - C_{11})\, p_1(y)$$

$$- (C_{1o} - C_{oo})\, p_o(y)\, P(H_o)\}\, dy \; .$$

This expression for $\bar{C}$ may be minimized by including in the region R_o only that portion of the y domain for which the integrand is negative. The region R_o where H_o is chosen is the region where

$$P(H_o)(C_{1o} - C_{oo})\, p_o(y) \geq [1 - P(H_o)](C_{o1} - C_{11})\, p_1(y)$$

If the likelihood ratio parameter λ is defined as

$$\lambda \triangleq \frac{p_1(y)}{p_o(y)} \; ,$$

then the decision rule is to choose H_1 if

$$\lambda \triangleq \frac{P(H_o)(C_{1o} - C_{oo})}{[1 - P(H_o)]\,(C_{o1} - C_{11})} \; .$$

If no cost is associated with a correct decision, and the errors of each kind are assigned equal cost, then

$$C_{oo} = C_{11} = 0$$

and

$$C_{o1} = C_{1o} = 1 \; .$$

Then, the decision rule is to choose H_1 if

$$\lambda \triangleq \left(\frac{p_1(y)}{p_o(y)}\right) \geq \left(\frac{P(H_o)}{1 - P(H_o)}\right).$$

This test is identical to the maximum a posteriori probability criterion.[4]

The above result is readily generalized to the case where the probability densities p_1 and p_o are functions of N-variates

$$p_1 = p_1(z_1, \ldots, z_N) \quad \text{and} \quad p_o = p_o(z_1, \ldots, z_N),$$

so that the decision rule is choose H_1 if

$$\lambda \triangleq \frac{p_1(z_1, \ldots, z_N)}{p_o(z_1, \ldots, z_N)} \geq \frac{P(H_o)}{1 - P(H_o)} .$$

For our case, we assume that it is equally likely that hypothesis H_1 or H_o is true and the decision then is choose H_1 if

$$\lambda \geq 1 .$$

This formulation represents the decision as a quotient of two multivariate PDF's. Thus, for the analyses discussed in this report, the requirement is to identify the form the hypothesis test assumes for each pair of multivariate densities that are being compared.

4.2 Specific Examples

In Section 3, the different multivariate probability densities which are used in the evaluation of the hypothesis testing procedure have been specified. The variates are all zero mean, unit variance, and uncorrelated. In the following sections the specific forms that the hypothesis test assumes for deciding between two alternative densities are given. Note that in all cases one of the two alternatives is the multivariate Gaussian density.

4.2.1 GAUSSIAN AND EXPONENTIAL HYPOTHESES

For this pair of densities, the likelihood ratio parameter is given by

$$\lambda = \frac{p_G}{p_E} = \left[\frac{2^{\frac{N+1}{2}} (2\pi)^{\frac{N-1}{2}} \Gamma\left(\frac{N+1}{2}\right)}{(2\pi)^{N/2} (N+1)^{N/2}}\right] \frac{\exp(-Q^2/2)}{\exp(-\sqrt{N+1}\, Q)}$$

$$= \left[\Gamma\left(\frac{N+1}{2}\right) e^{(N+1)/2}\right] \left[\sqrt{\pi}\left(\frac{N+1}{2}\right)^{N/2}\right]^{-1} \exp[-(Q - \sqrt{N+1})^2/2]$$

where

$$Q^2 = Z_1^2 + Z_2^2 + \dots + Z_N^2 \quad .$$

It is possible to rewrite the test in logarithmic form and assert that H_1 is true (the PDF is Gaussian) if $\ln \lambda \geq 0$. Then the result is: H_1 is true if

$$-(1/2)\,(Q - \sqrt{N+1})^2 \;\geq\; (1/2)\ln\pi - (1/2)(N+1) + \left(\frac{N}{2}\right)\ln\left[(N+1)/2\right] - \ln\left[\Gamma\left(\frac{N+1}{2}\right)\right]$$

or if

$$(\sqrt{N+1} - \sqrt{2B}) \;\leq\; Q \;\leq\; (\sqrt{N+1} + \sqrt{2B})$$

where

$$B = \ln\left\{\left[\Gamma\left(\frac{N+1}{2}\right) e^{\frac{N+1}{2}}\right]\left[\sqrt{\pi}\left(\frac{N+1}{2}\right)^{N/2}\right]^{-1}\right\} .$$

4.2.2 GAUSSIAN AND BESSEL HYPOTHESES

For this pair the likelihood ratio has the form

$$\lambda = p_G/p_K = \left(\frac{2^{\frac{N-1}{2}}\,\pi^{\frac{N+1}{2}}}{3^{N/2}\,2^{N/2}\,\pi^{N/2}}\right)\frac{(\sqrt{3}\sqrt{Q^2})^{\frac{N-3}{2}}\exp\{-Q^2/2\}}{K_{\frac{N-3}{2}}(\sqrt{3}\sqrt{Q^2})}$$

or

$$\lambda = \left(\frac{\sqrt{\pi}}{3^{N/2}\sqrt{2}}\right)\frac{(\sqrt{3}\sqrt{Q^2})^{\frac{N-3}{2}}\exp\{-Q^2/2\}}{K_{\frac{N-3}{2}}(\sqrt{3}\sqrt{Q^2})} .$$

Then the test is choose Gaussian if

$$0 \;\leq\; \ln(\pi/2) - \left(\frac{N+3}{2}\right)\ln 3 + \left(\frac{N-3}{2}\right)\ln Q^2 - Q^2 - 2\ln\left[K_{\frac{N-3}{2}}(\sqrt{3}\sqrt{Q^2})\right] .$$

4.2.3 GAUSSIAN AND LAPLACE HYPOTHESES

For this case, the likelihood parameter becomes

$$\lambda = p_G/p_L = \pi^{-N/2} \left(\frac{\exp[-(Z_1^2 + \ldots + Z_N^2)/2]}{\exp[-\sqrt{2}(|Z_1| + \ldots + |Z_N|)]} \right)$$

or

$$\lambda = (\pi^{-N/2} e^N) \exp\left[-(1/2) \sum_{i=1}^{N} (|Z_i| - \sqrt{2})^2 \right] .$$

Then, in terms of the log-likelihood ratio test we choose H_1 (Gaussian) if

$$2[N - (N/2) \ln \pi] \geq \sum_{i=1}^{N} (|z_i| - \sqrt{2})^2 .$$

4.2.4 GAUSSIAN AND GAMMA HYPOTHESES

For this pair of probability densities, the likelihood ratio parameter becomes

$$\lambda = p_G/p_\Gamma = \left(\frac{N^{\frac{N+3}{2}}}{\sqrt{2}\,\pi^{N/2}\,(N+2)^{\frac{N+1}{2}}} \right) \times$$

$$\times \left(\frac{\exp[-(Z_1^2 + \ldots + Z_N^2)/2]}{(|Z_1| + \ldots + |Z_N|) \exp[-\sqrt{2(N+2)/N}\,(|Z_1| + \ldots + |Z_N|)]} \right)$$

or

$$\lambda = \left(N^{\frac{N+3}{2}} e^{\frac{N+2}{N}} \right) \left(\sqrt{2}\,\pi^{N/2}\,(N+2)^{\frac{N+1}{2}} \right)^{-1} (|Z_1| + \ldots + |Z_N|)^{-1} \times$$

$$\times \exp\left[-(1/2) \sum_{i=1}^{N} (|Z_i| - \sqrt{2(N+2)/N})^2 \right] .$$

Then, for the log likelihood ratio test, the hypotheses test becomes: choose H_1 if

$$\left(\frac{N+3}{2}\right) \ln N - (1/2) \ln 2 - \left(\frac{N+1}{2}\right) \ln (N+2) - (N/2) \ln \pi + \left(\frac{N+2}{2}\right) \geq D$$

where

$$D \equiv \left[(1/2) \sum_{i=1}^{N} (Z_i - \sqrt{2 (N+2)/N})^2 + \ln \left(\sum_{i=1}^{N} Z_i \right) \right] .$$

4.3 Error Probabilities

In Section 1.2 and Section 4, we discussed error probabilities in relation to the decisions of the various hypothesis tests of interest. The concept is based on the identification of some statistic associated with both distributions which determines the two decision regions of the hypothesis test for the particular case. The errors involve probabilities of incorrect action when a given hypothesis is true (see Mood and Graybill).[7] For example, the probability of a type I error for decision D_o would be represented by the probability that the statistic value would fall in Region R_1 given that H_o is true and the corresponding type II error probability would be the probability that the value of the statistic would fall in in region R_o given that H_1 is true. The expressions for these conditional probabilities have been outlined in Section 4. 1. They reduce to integrations of the particular statistic PDF over the decision region of the opposite distribution.

As has been pointed out, it is not always clear as to what form the PDF of the statistic will have, particularly when the densities in the hypothesis test are complicated and dissimilar in form. For two cases, though, we have established the desired PDFs. Comparison of the analytic results for those cases with the Monte Carlo type results will help to assess the validity of the latter approach for cases where the analytic results are not available.

If we set the statistic, $q = \sum_{i=1}^{N} z_i^2$ and postulate that the multivariate distribution is Gaussian, then the density of the statistic is given by

$$p_q(q) = \begin{cases} [2^{N/2} \Gamma(N/2)]\, q^{\frac{N-2}{2}} e^{-q/2} & q \geq 0 \\ 0 & q < 0 \end{cases} .$$

For the case where the multivariate distribution is exponential, the density is

$$p_q(q) = \begin{cases} (N+1)^{N/2} \left[2\Gamma(N) \right]^{-1} q^{\frac{N-2}{2}} e^{-\sqrt{N+1}\sqrt{q}} & q \geq 0 \\ 0 & q < 0 \end{cases}$$

These relations are discussed in greater detail in Appendix B

5. RESULTS

In this section the topics discussed in the preceeding sections have been combined. Computer programs were written for generating the different sets of random variates and using them in the associated simple alternative hypothesis tests. These results, the analysis of some of the related questions regarding the numerical aspects, and the analytic error results will all be presented.

5.1 Monte Carlo Applications

Five-variate jointly distributed sets were generated for the various forms. In each case the alternative decision was that the variates came from a Gaussian distribution. For the appropriate hypothesis test for such cases, the total of correct decisions was obtained and the corresponding number of correct decisions was then determined for an equivalent number of five-variate Gaussian samples. The probability for correct Gaussian decisions will change from case to case since the form of the test is dependent on the two distributions considered.

The first comparison is between the Gaussian and the multivariate exponential. The Gaussian distributions were correctly identified 72.2 percent of the time. Alternatively, the exponential sets of variates were correctly chosen in 43.3 percent of the cases. Next, we looked at decisions between Gaussian and other non-independent densities. The test decided correctly for the Bessel variates at a 50 percent rate and for the related Gaussian sets at 76 percent. For gamma-related variates, 56 percent correct decisions were made, while for that case the alternative Gaussian decision was correct 64.7 percent of the time. Finally, we considered an alternative hypothesis test that consisted of sets of Laplacian variates. This is a case where the variates, like Gaussians, are independent. This was included to see if the single-function integration required for an independent joint distribution would affect the ratio of correct decisions. For the Laplace case, the test was correct 56.5 percent of the time and correspondingly 68.5 percent for the Gaussians.

5.2 Effect of Numerical Integration

As was discussed in Section 2 and Section 3, independent variates require only a single function to be integrated and for many cases the integration does not have to be done numerically. In contrast, the nonindependent variates require successive integrations of different functions that include the preceding variates of the set. Thus, there is a question as to whether these repeated numerical integrations could have an effect on the values assigned to the variates and consequently, on the outcome of the decision process for those variates.

To address the question, the gamma-like set of jointly distributed variates was examined again. For the second analysis, the forms for the five successive marginal densities of the gamma-like function can be analytically integrated to give the corresponding forms for their respective cumulative distribution functions. These forms can then be used directly to relate to the set of uniform variates. Thus, we can compare the results of the hypothesis test for the same dependent multivariate distribution when the numerical integrations have been removed from the process. For the five-variate gamma-like case without numerical integration, the test correctly identified the sets in 58.5 percent of the cases and for the related Gaussian test the result was correct 66 percent of the time.

5.3 Effect of an Increase in Variates

In parameter estimation theory we can obtain a better estimate for the parameters of a distribution by increasing the number of samples of the population in the estimator. Similarly, in typical hypothesis testing, the results would tend to improve as we increased the number of observations that appear in the product of the likelihood functions that determine the likelihood ratio for the hypothesis test. However, that is not what we are constructing here. In the present formulation we employ only a single observation of a multivariate PDF as the complete likelihood ratio. For this to be a similar situation to the preceding ones, we would have to use multiple observations of the joint density population. Increasing the number of variates in the respective joint distributions is not equivalent to increasing the observations; thus, it is not obvious what the effect such an increase would have on the reliability of the hypothesis testing procedure. This aspect has been studied for the two cases of multivariate exponential and Bessel distributions.

As can be seen in Table 1 the two forms result in different decision histories as the number of variates increases. The first part of the table shows the results for each multivariate level when two data sets, one Gaussian and one exponential, are entered into the Gaussian/exponential hypothesis tests. The second part of the table shows the corresponding results for two additional sets of populations, one Gaussian and the other Bessel, in the Gaussian/Bessel Hypothesis tests. In the

evaluation, the large number of Gaussian sets are tested against the decision criterion. Then a second series of either Bessel variates or exponential variates is tested. The ratios of decisions correctly identifying the known population, compared to the total tries for that particular population, determine the percentages shown in the tables. For both sets of hypothesis tests, the percentage of correct Gaussian decisions increases with N as does the percentage of correct Bessel distribution decisions. This is not true for the exponential populations in the Gaussian/exponential tests.

Table 1. Percentage of Correct Decisions for Four Sets of RNG Populations in Gaussian/Exponential and Gaussian/Bessel Tests as a Function of the Number of Variates

Gaussian and Exponential Sets		
N	Correct Gauss.	Correct Exp.
5	72.1	43.5
10	74.3	43.4
25	74.5	41.5
50	75	40.5
Gaussian and Bessel Sets		
N	Correct Gauss.	Correct Bessel
5	75	50
10	83.7	54
25	86	62.5
50	93.5	70

5.4 Analytic Results and Comparisons

The corresponding analytic error-evaluation for the test of Gaussian or exponential hypotheses has been discussed in Section 1 and Section 4. The reason for comparing the analytic and random variate generation approaches is to establish confidence in the Monte Carlo approach which can be applied more generally, especially in cases where the analysis is too complicated. Here we will present the results for both types of error with each hypothesis and compare the results with the related Monte Carlo percentages.

For N = 5, the type I error associated with the exponential hypothesis is based on P_E $(2.51 \leq Q^2 \leq 10.987)$ and the related type II error is based on the probability

as a function of the number of variates as determined analytically. For completeness, the table includes the equivalent results for the Bessel case. This aspect required specific determination by computer of a relationship between the decision regions and the quantity, Q^2. The functions are much more complicated in this case and it should be noted that the odd numbered variates were used, based on the need for available forms of the Struve function. The analytic trends appear quite similar to those observed for the random number generation cases.

5.5 Summary of Results

In the preceding sections a number of interesting results have been enumerated. In this section we will attempt to place those results in perspective.

In all the simple alternative cases studied, the Gaussian correct percentages are always the higher of the two. Also, there does not appear to be any significant distinction between independent and nonindependent forms as the alternative.

For the case where numerical integration of a nonindependent multivariate form was replaced by an equivalent analytic result, the change was slight. Thus, this does not appear to be a problem in testing the error probabilities to be expected for a decision.

The decrease in correct decisions with increased variates in the exponential case contrasts to the Gaussian results for that pair of alternatives and to both sets of results when the Bessel form is the alternative to the Gaussian. The difference in behavior appears in both the random variate and the analytic results.

In the exponential case, a further test was made for N = 100. As a result of the prolonged run-time on the computer this case was halted after 200 decisions. At that point, the ratio of correct decisions was: Gaussian 79 percent and exponential 38.5 percent. This is slightly off the predicted trend line value for the respective cases but still is quite close. The corresponding analytic results for that case are: Gaussian 75.7 percent and exponential 40.8 percent.

For all the levels of variates considered there is good agreement between the analytic and random number generated results for both exponential and Bessel alternatives. This would tend to give confidence in the use of the computer technique to evaluate the decision process.

In all these results the main factor is that the decision process and the associated decision regions are based on defining those portions of the probability space where one of the two alternatives has a greater probability of occurring. As the test now stands, there is no assessment in the decision as to the relative magnitude of the second distribution in that region. The high probability of incorrect decisions for the alternative distributions is related to such variates having a high

probability in regions where, nevertheless, the Gaussian PDF dominates. Since we are not dealing with multiple observations of these multivariate distributions, we are limited in our ability to control the resultant unsatisfactorily large error probabilities. It should be noted that for the Bessel case the error is decreasing as the number of variates increases, which is a desirable result in terms of applying these results to terrain characterization.

6. TERRAIN CHARACTERIZATION APPLICATIONS

The purpose of the statistical studies described in this report is to establish confidence levels and limitations for the hypothesis testing procedures used in the terrain characterization studies of our program to analyze electromagnetic scattering from rough surfaces. In previous sections we have pointed out that we are using sets of uncorrelated variates having distributions with zero mean and unit variance. In this section we will be showing the relation between those conditions and the more general ones required in terrain studies.

6.1 Nonzero Mean, Correlated Variates

In the terrain analysis hypothesis testing using multivariate density functions, the decisions are based on the values of the quadratic form $F = (\underset{\sim}{z}' - \underset{\sim}{\mu})^T \underset{\sim}{R}^{-1} (\underset{\sim}{z}' - \underset{\sim}{\mu})$. This involves the inverse of the covariance matrix and the associated nonzero mean (μ), correlated heights, $\underset{\sim}{z}'$. In our earlier report[1] we discussed the transformations of coordinates used to generate appropriate PDFs for the variates. Similar relations will be used here to show that it is sufficient to derive relationships for sets of uncorrelated, zero mean variates. The results then apply to the general multivariate case as well.

First, consider the general multivariate Gaussian PDF. It has a form that is well known.[7]

Let $\underset{\sim}{z}' = (z_1', z_2', \ldots, z_N')$ be an N-dimensional random variable in vector form. Then, the multivariate Gaussian PDF is given by

$$p_G(z_1', \ldots, z_N') = \left[|\underset{\sim}{R}|^{1/2} (2\pi)^{N/2} \right]^{-1} \exp\{-Q^2/2\}$$

where

$$Q^2 = [(\underset{\sim}{z}' - \underset{\sim}{\mu})^T \underset{\sim}{R}^{-1} (\underset{\sim}{z}' - \underset{\sim}{\mu})]$$

where T [illegible] transpose, $\underset{\sim}{R}$ is a [illegible] matrix [illegible] elements [illegible] constants and [illegible] [illegible] the case considered in this report [illegible] matrix [illegible] matrix for the variates, where

$$R_{ij} = \langle (z_i' - \mu_i)(z_j' - \mu_j) \rangle$$

and $\langle\rangle$ denotes expectation value. Note that the PDF contains the quadratic form $F = (\underset{\sim}{z}' - \underset{\sim}{\mu})^T \underset{\sim}{R}^{-1} (\underset{\sim}{z}' - \underset{\sim}{\mu}) = Q^2$.

The question to be resolved is the relation between the very general form represented by this relation and the more simplified forms used in the discussions of this report. Trivially, for $\underset{\sim}{y} = \underset{\sim}{z}' - \underset{\sim}{\mu}$, $F = \underset{\sim}{y}^T \underset{\sim}{R}^{-1} \underset{\sim}{y}$. Next, consider the eigenvalues of the covariance matrix, $\underset{\sim}{R}$ and their associated eigenvectors; let $\underset{\sim}{E}$ be the eigenvector column matrix for $\underset{\sim}{R}$ and $\underset{\sim}{\lambda}$ the eigenvalue vector. Then, consider the matrix

$$\underset{\sim}{A} = \begin{pmatrix} e_{11} & \cdots & e_{N1} \\ \vdots & & \vdots \\ e_{1N} & \cdots & e_{NN} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\lambda_N} \end{pmatrix}$$

and

$$\underset{\sim}{A}^{-1} = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\lambda_N} \end{pmatrix}^{-1} \begin{pmatrix} e_{11} & \cdots & e_{1N} \\ \vdots & & \vdots \\ e_{N1} & \cdots & e_{NN} \end{pmatrix} .$$

The vector $\underset{\sim}{u}$, defined by $\underset{\sim}{u} = \underset{\sim}{A}^{-1} \underset{\sim}{y}$, is a new vector representation consisting of uncorrelated, $\sigma^2 = 1$ variates. Then,

$$\underset{\sim}{y} = \underset{\sim}{A}\underset{\sim}{u} \qquad \text{and} \qquad \underset{\sim}{y}^T = \underset{\sim}{u}^T \underset{\sim}{A}^T .$$

Also,

$$\underset{\sim}{R} = \underset{\sim}{A}\,\underset{\sim}{A}^T \qquad \text{and} \qquad \underset{\sim}{R}^{-1} = (\underset{\sim}{A}^T)^{-1} \underset{\sim}{A}^{-1} .$$

Then,

$$F = \underset{\sim}{y}^T \underset{\sim}{R}^{-1} \underset{\sim}{y} = \underset{\sim}{u}^T \underset{\sim}{A}^T (\underset{\sim}{A}^T)^{-1} \underset{\sim}{A}^{-1} \underset{\sim}{A} \underset{\sim}{u} = \underset{\sim}{u}^T \underset{\sim}{u}.$$

Thus, for our terrain hypothesis testing, the value of the quadratic form is unaffected by the degree of covariance among the variates and the results obtained using sequences of uncorrelated, zero mean, unit variance variates are quite general.

6.2 Discussion

As has been pointed out, the hypothesis test is based on minimizing the total possibility of error. Further, we assigned equal probability for the heights to be exponential or Gaussian and equal costs to either incorrect decision. Given these factors, we found that the test was more likely to identify correctly a sample from a Gaussian distribution than from an exponential one. Alternatively, we could abandon the minimal cost and modify the form of the test by readjusting the decision regions to allow equal error probabilities. This is equivalent to a minimal cost condition, where we have biased the cost of an incorrect decision for one distribution or the a priori probabilities. For instance, we could use terrain features to assign a value greater than one half to the probability that points are distributed in a given form. These aspects have not yet been incorporated into the decisions of the terrain characterization program.

The preceding conclusion is interesting in terms of the decisions generated by the terrain data base in Massachusetts used in the initial electromagnetic calculations. In the original report,[1] there is some discussion of those results. When corrections to eliminate some of the errors induced by inverting a large-covariance matrix are included, the results are still overwhelmingly in favor of the points being from an exponential distribution. This is disturbing in light of the expected bias towards Gaussian distributions found in the error analysis. Consequently, there are several additional approaches to characterization currently being pursued.[2] There still appear to be numerical problems associated with the decision process that will have to be resolved. One additional aspect of interest is the effect that the quantization in height of the original data base has on the resultant distributions; this also is under investigation.

In this report, we have discussed some of the statistical implications of the hypothesis testing approach to terrain characterization as applied to our electromagnetic scattering analyses. The emphasis has been on understanding constraints, confidence levels, and errors associated with the decision process. This

leads to the development of random number generation techniques for multivariate nonindependent distributions and to further consideration of the effects of numerical accuracy and errors on the decision outcome.

References

1. Lennon, J.F. and Papa, R.J. (1980) Statistical Characterization of Rough Terrain, RADC-TR-80-9, RADC/EE Hanscom AFB, MA.
2. Papa, R.J., Lennon, J.F., and Taylor, R.L. (1980) Electromagnetic Wave Scattering From Rough Terrain, RADC-TR-80-300, RADC/EE Hanscom AFB, MA.
3. Jenkins, G.M., and Watts, D.G. (1968) Spectral Analysis and Its Applications, Holden-Day, San Francisco, CA.
4. Whalen, A.D. (1971) Detection of Signals in Noise, Academic Press, New York, NY.
5. Shreider, Yu., A. (1966) The Monte Carlo Method, Pergamon Press, Oxford.
6. Papoulis, A. (1965) Probability, Random Variables and Stochastic Processes, McGraw-Hill, New York, NY.
7. Mood, A.M., and Graybill, F.A. (1963) Introduction to the Theory of Statistics, McGraw-Hill, New York, NY.
8. Lennon, J.F. (1980) The Derivation of a Multivariate Probability Density Function Having an Exponential-Type Bivariate Marginal Density, RADC-TR-80-153, RADC/EE, Hanscom AFB, MA.
9. Morrison, D.F. (1976) Multivariate Statistical Methods, McGraw-Hill, New York, NY.

Appendix A

A Multivariate Probability Distribution Function: Derivation and Analysis

In this study the cumulative distribution functions for the cases of independent variates are analytically integrable and hence the relation between the uniform random variates and the new set is readily obtained. For the nonindependent cases of exponential or Bessel PDFs, however, this is not so and the value of the new random variate has to be determined as the result of a numerical integration. To examine whether this introduces any effect into the results, particularly those of the hypothesis testing procedures, we formed a new type of nonindependent PDF. This form is analytically integrable for the case N = 5 and hence allows a comparison between results when the same set of five functions are evaluated both analytically and numerically.

The form selected for this purpose is

$$p(z_1, \ldots, z_N) = C_1(|z_1| + |z_2| + \ldots + |z_N|) \exp[-C_2(|z_1| + \ldots + |z_N|)] .$$

As in previous work we use the zeroth and second moment integrals to determine the normalization constants to satisfy the requirements for this to be a probability density. This leads to

$$C_1 = C_2^{N+1} / (2^N N) \qquad \text{and} \qquad C_2^2 = 2(N+2)/N .$$

In addition to the parameters of the general multivariate density function, the procedures also require the form for the general marginal density of arbitrary order,

$$p(z_1, \ldots, z_L) = \left(\frac{2^{N-L} C_1}{C_2^{N-L+1}}\right) \left((N-L) + C_2[\,|z_1| + \ldots + |z_L|\,]\right) \times$$

$$\times \exp[-C_2(|z_1| + \ldots + |z_L|)] \qquad \text{for } L \leq N.$$

For actual use we employed N = 5 and the appropriate set of functions is:

$$p(z_1, \ldots, z_5) = C_1(|z_1| + \ldots + |z_5|) \exp[-C_2(|z_1| + \ldots + |z_5|)],$$

$$p(z_1, \ldots, z_4) = [2C_1/C_2^2][1 + C_2(|z_1| + \ldots + |z_4|)] \times$$

$$\times \exp[-C_2(|z_1| + \ldots + |z_4|)] ,$$

$$p(z_1, z_2, z_3) = (4C_1C_2^3)[2 + C_2(|z_1| + |z_2| + |z_3|)] \times$$

$$\times \exp[-C_2(|z_1| + |z_2| + |z_3|)] ,$$

$$p(z_1, z_2) = (8C_1/C_2^4)[3 + C_2(|z_1| + |z_2|)] \exp[-C_2(|z_1| + |z_2|)] ,$$

and

$$p(z_1) = (16C_1/C_2^5)[4 + C_2|z_1|] \exp[-C_2|z_1|] ,$$

where

$$C_1 = [14^3/(5^4\, 2^5)]$$

and

$$C_2^2 = 14/5 .$$

Appendix B

Functions of a Sequence of Random Variates

In this appendix we will discuss the appropriate forms for the probability density function for the statistic $q = \sum_{i=1}^{N} z_i^2$ when the original multivariate density is either Gaussian or exponential with zero mean and unit variance.

The derivation is based on the principle that if there is a collection of random variates $\{z_i\}$ and a function $g(z_1, z_2, \ldots, z_N)$ then generally, $\chi = g(z_1, \ldots, z_N)$ is a random variate and we are concerned with determining the form of its density function and that of additional functions, $g_i(\chi)$, once the basic form is established. Consider the hypershell ΔD_χ of the $(z_1, \ldots, z_N)$ space such that

$$\chi < \sqrt{z_1^2 + \ldots + z_N^2} < \chi + d\chi .$$

Here $\chi \equiv \sqrt{z_1^2 + \ldots + z_N^2}$ is the basic function of the random variates and $q = \chi^2$ is the secondary function of interest in the error probabilities. The volume of the hypershell is given by

$$dV = \left[2^N \pi^{\frac{N-1}{2}} \Gamma\left(\frac{N+1}{2}\right)/\Gamma(N)\right] \chi^{N-1} \, d\chi .$$

Papoulis[16] shows how the two results $p_\chi(\chi)$ and $p_q(q)$ are determined for the Gaussian multivariate case,

$$p(z_1, \ldots, z_N) = (2\pi)^{-N/2} \exp\left[-\tfrac{1}{2}(z_1^2 + \ldots + z_N^2)\right] .$$

The final result for $q = \chi^2$ is

$$p_q(q) = \begin{cases} [2^{N/2}\,\Gamma(N/2)]^{-1}\, q^{\frac{N-2}{2}} \exp[-q/2] & q \geq 0 \\ 0 & q < 0 \end{cases} .$$

Similarly, for the exponential case where

$$p(z_1, \ldots, z_N) = (N+1)^{N/2}\left[2^{\frac{N+1}{2}}\,(2\pi)^{\frac{N-1}{2}}\,\Gamma\left(\frac{N+1}{2}\right)\right]^{-1} \times$$

$$\times \exp\left[-\sqrt{N+1}\,\sqrt{z_1^2 + \ldots + z_N^2}\right] ,$$

we obtain

$$p_\chi(\chi) = \begin{cases} [(N+1)^{N/2}/\Gamma(N)]\,\chi^{N-1} \exp[-\sqrt{N-1}\,\chi] & \chi \geq 0 \\ 0 & \chi < 0 \end{cases}$$

Then

$$p_q(q) = \begin{cases} [(N+1)^{N/2}/(2\Gamma(N))]\, q^{\frac{N-2}{2}} \exp[-\sqrt{N+1}\,\sqrt{q}\,] & q \geq 0 \\ 0 & q < 0 \end{cases} .$$

MISSION

of

Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of technical competence is provided to ESD Program Offices (POs) and other ESD elements. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.

Printed by
United States Air Force
Hanscom AFB, Mass. 01731

DATE
ILMED
-8